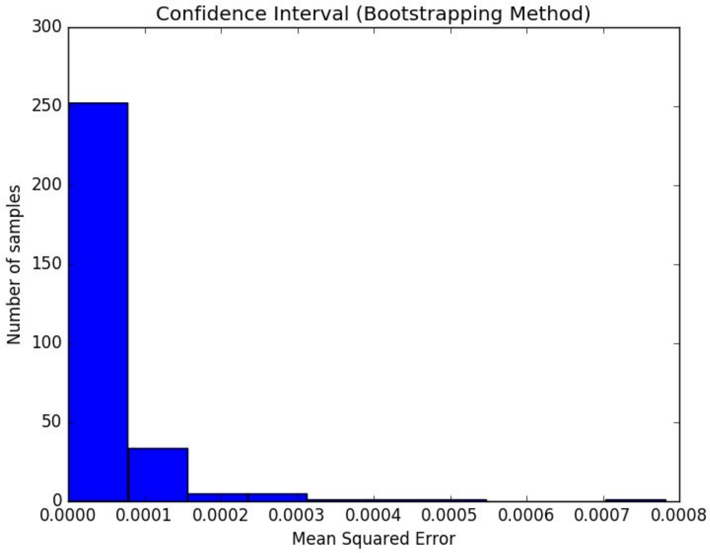
Armand Nasseri
912679383
ECS 171

Homework 3

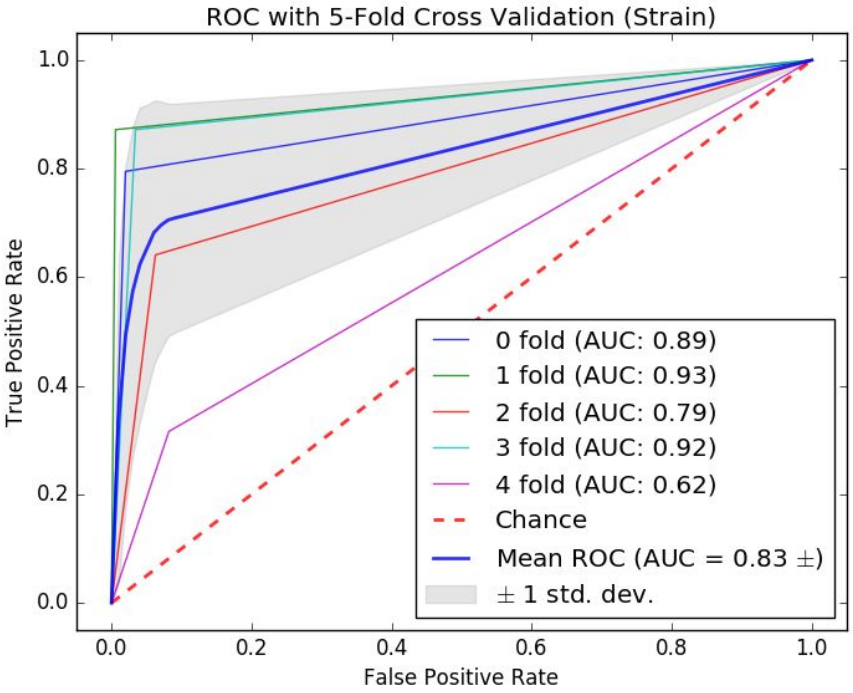
For this homework assignment, I have created 7 different files that correspond to each of the problems in the assignment description. Each file will have to be run separately to view the output. Note: some files will take longer than others due to the nature of SVM's efficiency and the size of the dataset.

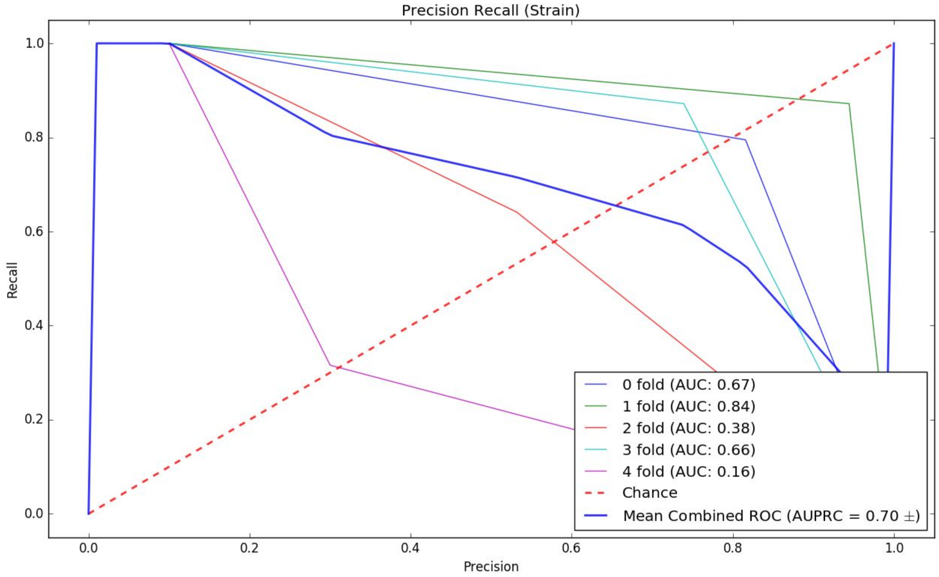
1) The regularized regression technique that I chose was **lasso regression.** From sklearn, lasso regression is defined as (1 / 2 * $n_{samples}$)) * | | y - Xw $||_2^2$ + alpha * | | w $||_1$. Lasso regression is used to reduce the amount of variables to create a more accurate model. To find the optimal amount of features to reduce from lasso regression, I needed to determine the optimal alpha to use in the above equation. I performed a grid search on several alpha values and it was determined that **alpha = 0.0001** was the most optimal. By applying 5-fold cross validation and using this alpha value, the amount of features with non-zero coefficients was **166**. The generalization error with these given parameters was **0.03167**.
2) For finding the confidence interval, I used the bootstrapping method. For configuration, I had a total of **300 iterations** and used a train-test split for dataset samplings. I calculated the mean of the dataset and the predictor to be used for each iteration in calculating the mean squared error. It is to be assumed that mean expression value is the mean of the growth rate. I then fit my lasso regression model with the optimal alpha value of **0.0001** on the training samples. The score per iteration was then placed into a stats list in which the final calculation of the confidence interval could take place.
3) The confidence interval of 95% of predicted growth for a bacterium whose genes are expressed exactly at the mean expression value is between **0.00001 and 0.02435.** Below is a chart to illustrate this:



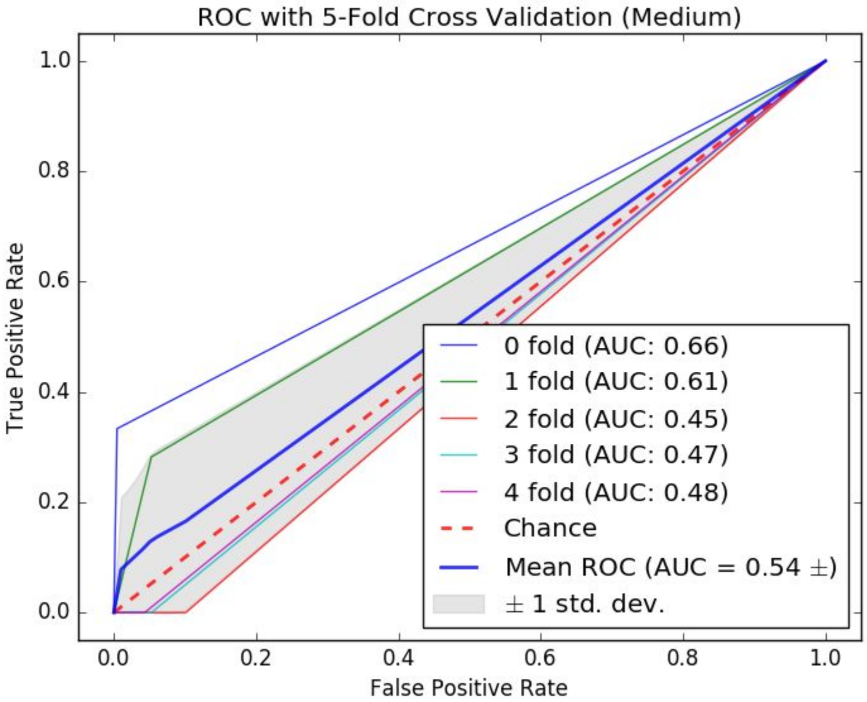
4) For this problem, I used the non-zero weighted features from lasso regression from problem 1. The number of non-zero features varied from Strain, Medium, Stress, and Gene Perturbation. Below are the ROC and PR curves with the corresponding number of features in each curve after applying 5-fold cross validation.

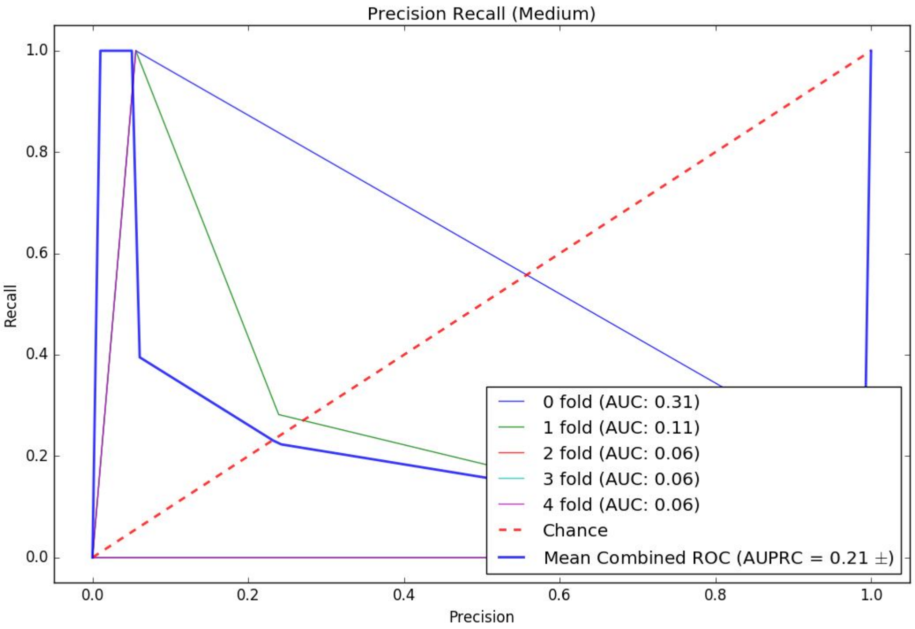
Number of Features for Strain: 171



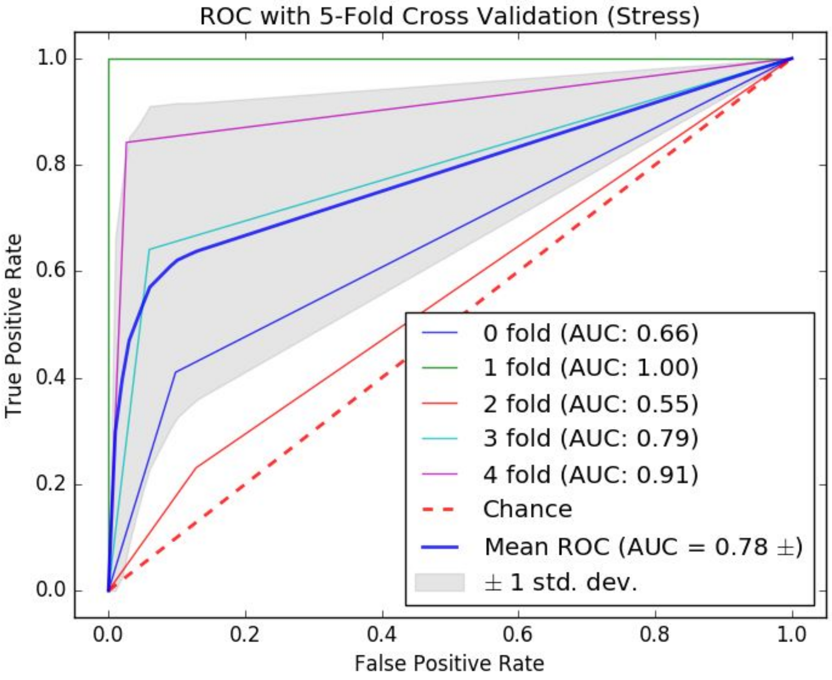


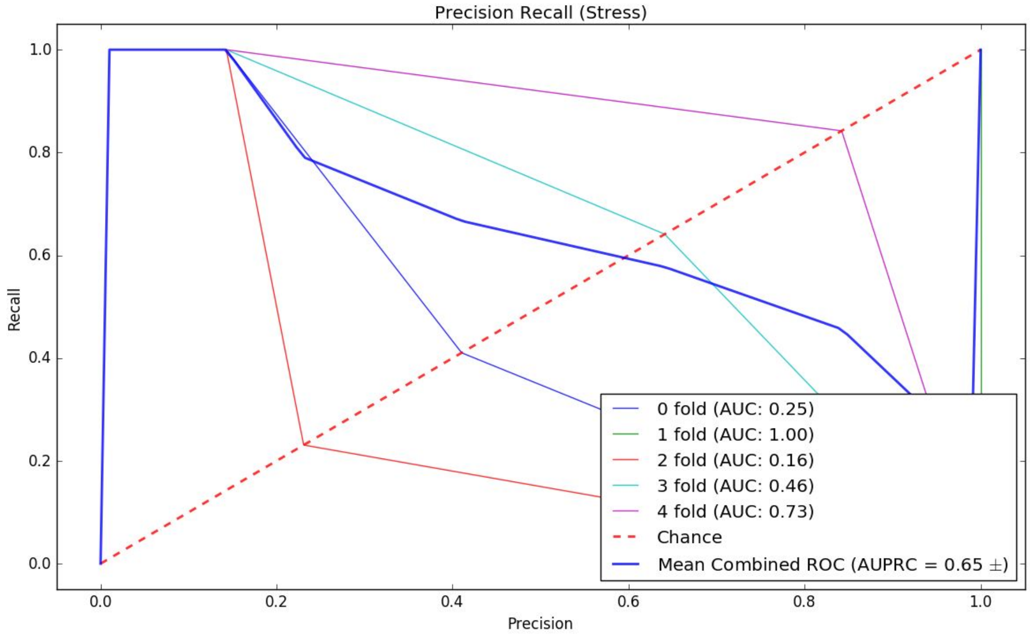
Number of Features for Medium: 26



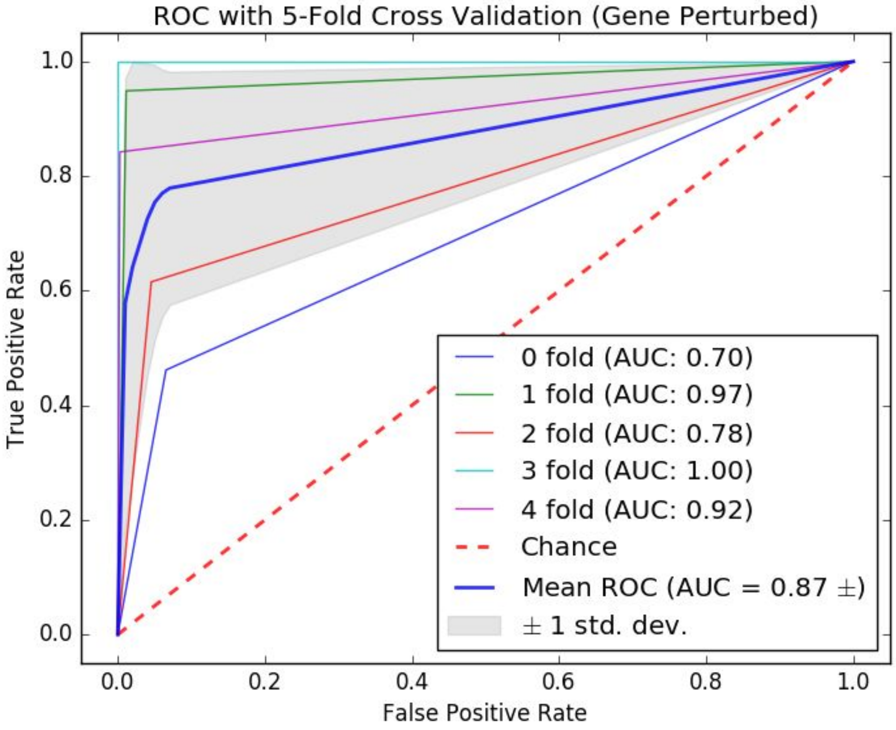


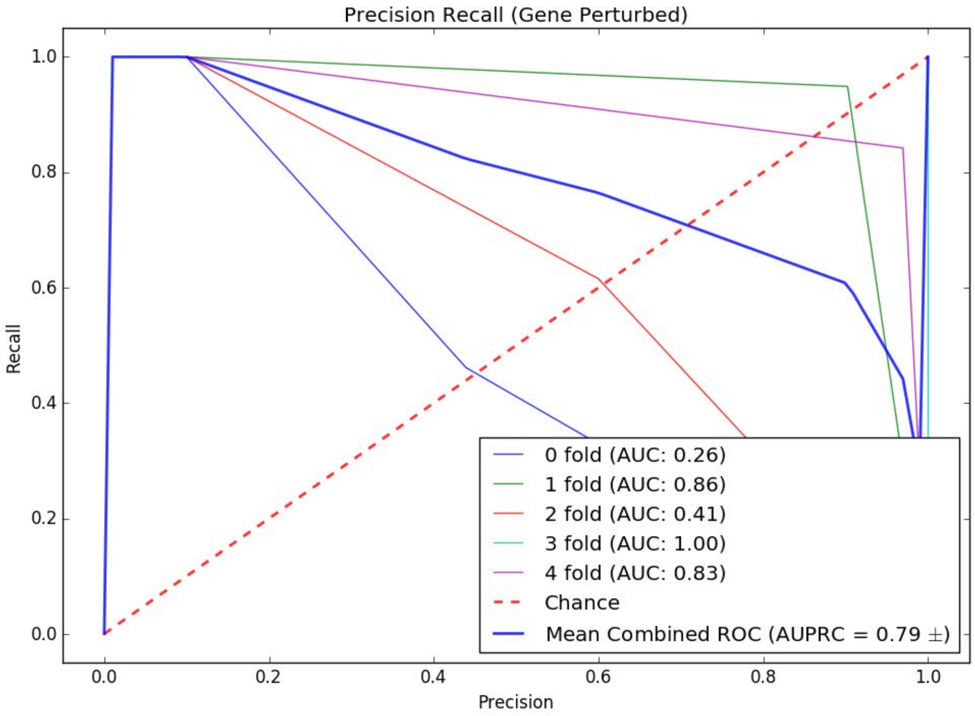
Number of Features for Stress: 1





Number of Features for Gene Perturbation: 12





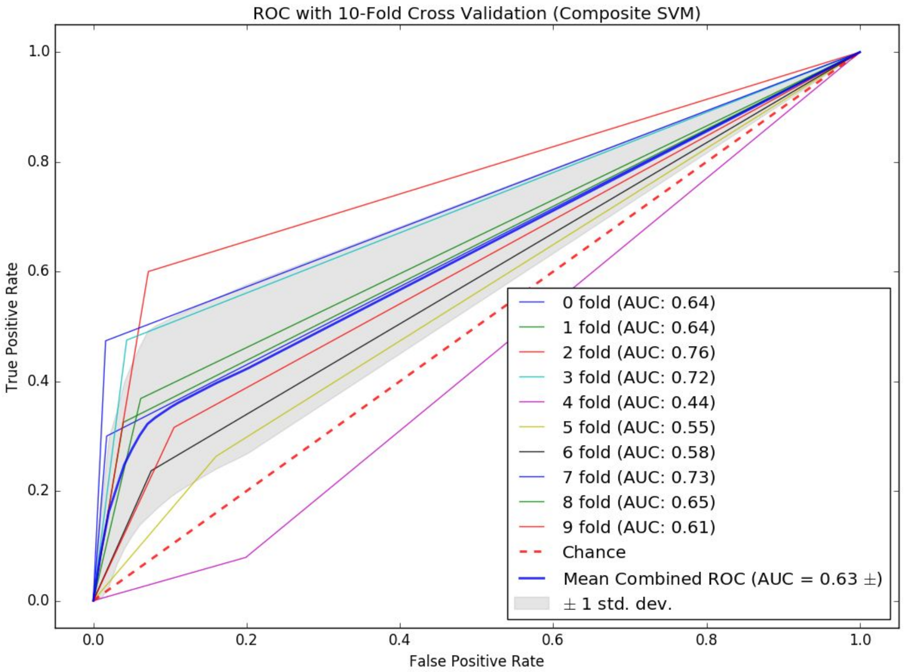
Averages for AUC values

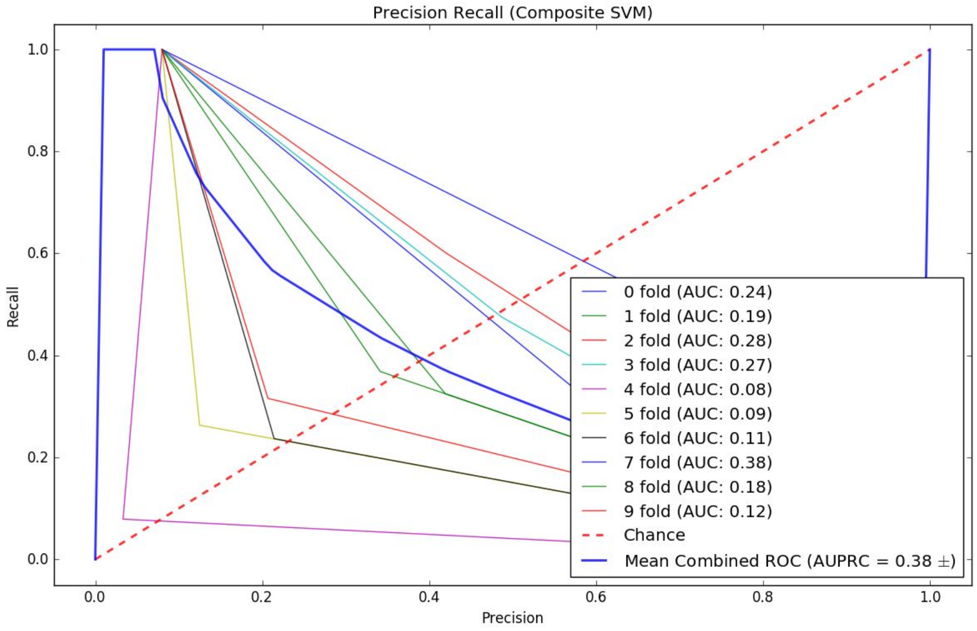
Strain	Medium	Stress	Gene Perturbed
0.83	0.54	0.78	0.87

Averages for AUPRC values

Strain	Medium	Stress	Gene Perturbed
0.70	0.21	0.65	0.79

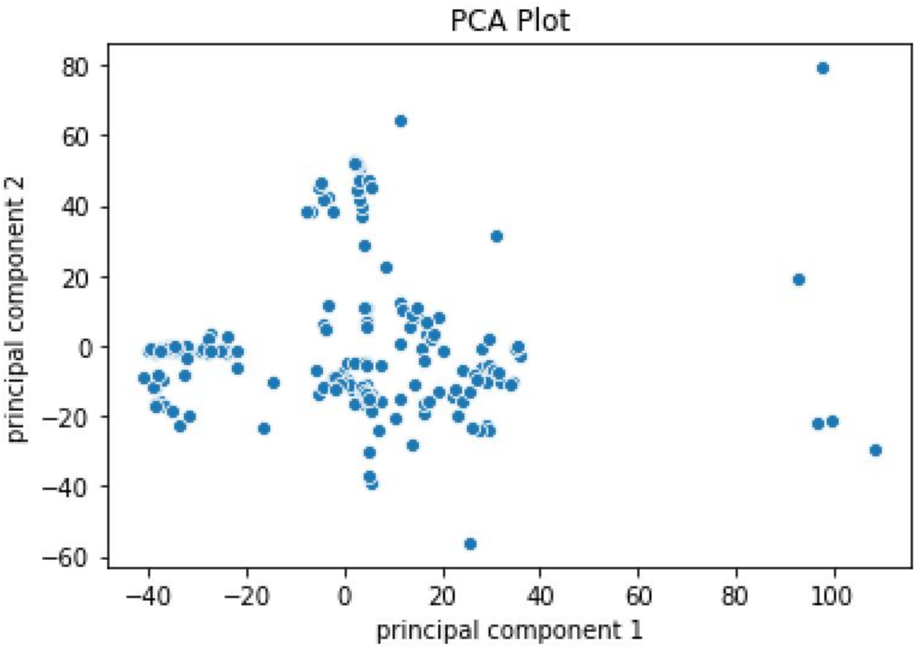
5) Below are the composite SVM 10-fold cross-validation AUC/AUPRC plots. It appears that this classifier performs worse than the two individual classifiers because both the AUC/AUPRC values decreased. It would seem that we are better off building two separate classifiers to simultaneously predict these two features. The baseline prediction performance (null hypothesis) is 50% as shown by the “Chance” line.



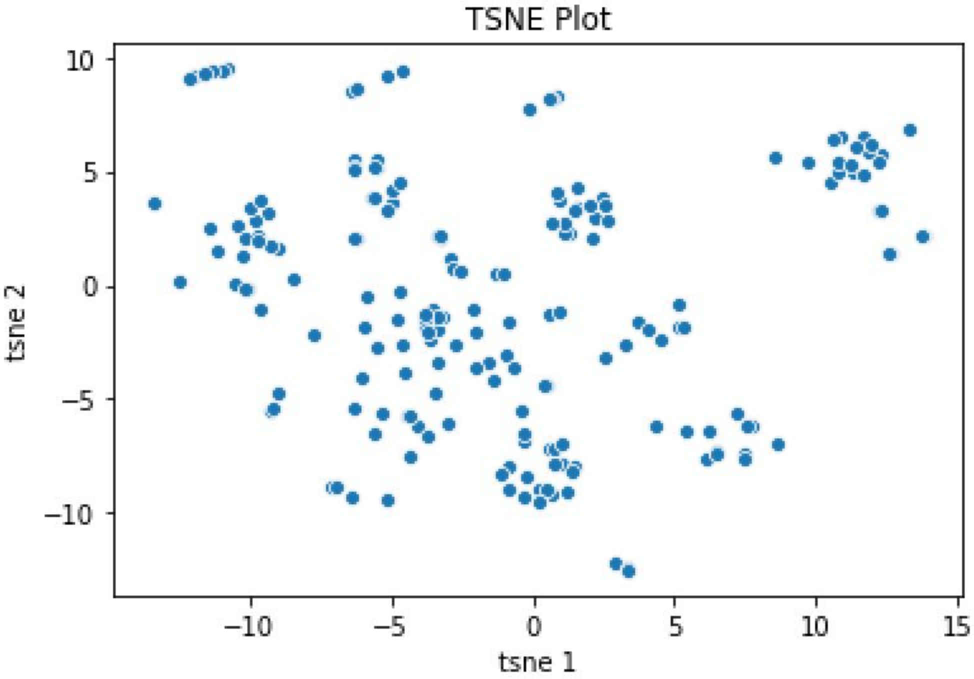


6) Gene Expression plots listed below:

PCA PLOT



TSNE PLOT



7) For this problem, I applied PCA and TSNE along with a 10-fold cross validation to obtain the following graphs. Below are the ROC and PR curves along with a table set of the 10-fold cross validation AUC/AUPRC values. Based on the results, it seems that T-SNE is the best preprocessing approach.

Averages for AUC values (PCA)

Strain	Medium	Stress	Gene Perturbed
0.82	0.50	0.78	0.89

Averages for AUPRC values (PCA)

Strain	Medium	Stress	Gene Perturbed
0.72	0.25	0.65	0.80

Averages for AUC values (TSNE)

Strain	Medium	Stress	Gene Perturbed
0.84	0.50	0.78	0.89

Averages for AUPRC values (TSNE)

Strain	Medium	Stress	Gene Perturbed
0.72	0.25	0.65	0.80